한국통계학회논문집 제1권1호
1994년 12월 pp. 33-40

Robust Estimation and Outlier Detection[1)]

Myung Geun Kim[2)]

Abstract

The conditional expectation of a random variable in a multivariate normal random vector is a multiple linear regression on its predecessors. Using this fact, the least median of squares estimation method developed in a multiple linear regression is adapted to a multivariate data to identify influential observations. The resulting method clearly detect outliers and it avoids the masking effect.

1. Introduction

In a multivariate data, the sample mean vector and the sample covariance matrix are very sensitive to outliers, and the latter is more sensitive than the former. This means that any analysis based on them can be exposed to outliers. For example, the well-known Wilks' statistic (1963) based on the usual Mahalanobis distance for detecting outliers suffers from the masking effect caused by insufficient detection of closely located outliers so that multiple outliers do not necessarily have a large distance. The masking phenomenon is that less extreme outliers mask the effect of the most extreme observations under investigation as outliers.

One way to remedy the masking effect is to use robust estimators. Substituting robust estimators for the maximum likelihood estimators in computing the Mahalanobis distance may mitigate such a masking effect. A global measure of robustness is the breakdown point whose finite-sample version is defined as the minimum fraction of outliers that will spoil the estimator completely (Rousseeuw and Leroy, 1987, p.10).

Campbell (1980) proposed to replace the maximum likelihood estimators for mean vector and covariance matrix by the respective M estimator to compute robust Mahalanobis distance and applied the result to outlier detection and principal component analysis.

But the breakdown point of M estimators is at most $1/(p+1)$ (Rousseeuw and Leroy, 1987,

1) This paper was supported in part by NON-DIRECTED RESEARCH FUND, Korea Research Foundation

2) Department of Applied Statistics, Seowon University, 231 Mochung-Dong, Chongju, Chung-Buk, 360-742, KOREA

p.253), where p is the number of variables in a random vector. As the number of variables increases, the breakdown point goes down to zero. Thus the development of outlier detection based on estimators with high breakdown points is necessary. Rousseeuw and van Zomeren (1990) introduced the minimum volume ellipsoid method for obtaining highly robust estimators and applied it to outlier detection. For mean vector the center of the minimum volume ellipsoid covering half of the observations is taken, and the minimum volume ellipsoid multiplied by a correction factor is taken as an estimator for covariance matrix. The breakdown point of the minimum volume ellipsoid estimator is nearly 50%. The minimum volume ellipsoid estimators are obtained using the resampling algorithm (Rousseeuw and van Zomeren, 1990, p.638) by drawing subsamples of size p+1, and sometimes the covariance matrices for some subsamples are nearly singular. Both methods mentioned above lack of a formal test for outliers because the distributions of the corresponding statistics are not available.

The least median of squares method has been introduced by Rousseeuw (1984) for detecting multiple outliers in the context of regression analysis. The least median of squares estimator for the regression parameter is found by minimizing the median of squares of the residuals and its breakdown point is 50%(Rousseeuw and Leroy, 1987, p.118). The least median of squares method is powerful in detecting multiple outliers and avoids the masking effect (see Rousseeuw and Leroy, 1987, Chap. 3).

It is well known that the conditional expectation of a variable in a multivariate normal vector is a multiple linear regression on its given predecessors (Johnson and Wichern, 1992, p.332). In this work the least median of squares method is adapted to a multivariate normal data to screen out possible outliers and the remaining data will be used for getting robust estimators for mean vector and covariance matrix. The resulting procedure applied to outlier detection is very effective in avoiding the masking effect. Numerical examples are given for illustration.

2. A procedure for detecting outliers

Let $\mathbf{x}_r = (\mathbf{x}_{1r}, \cdots, \mathbf{x}_{pr})^T (\mathbf{r} = 1, \cdots, \mathbf{n})$ be a random sample from a p-variate normal distribution with mean vector $\mu = (\mu_1, \cdots, \mu_p)^T$ and positive-definite covariance matrix Σ. We will not notationally discriminate between random vector and its realization because it will be clear from the context.

Some statistics for detecting outliers are based on the following Wilks' statistics (1963)

$$\{(x_i-m)^T W^{-1}(x_i-m)\}^{1/2} \quad (1)$$

computed for each observation (for example, Bacon-Shone and Fung, 1987, Caroni and Prescott, 1990), where **m** is the sample mean vector and **W** is the sample covariance matrix. The statistic (1) can be derived as the likelihood ratio statistic for the mean slippage model. Observations having large Mahalanobis distance are regarded as outliers. This procedure performs well if only a single outlier is present, but otherwise suffers from the masking effect (for more details, refer to Barnett and Lewis, 1984, Chap. 9).

The conditional expectation of a variable in a multibariate normal vector is a multiple linear regression on its given predecessors. Thus the multiple linear regression of the ith random variable on its given predecessors for each i = 1,...,p can be expressed as

$$\mathrm{x}_{ir} = \beta_{i0} + \beta_{i1}\mathrm{x}_{ir} + \cdots + \beta_{i,i-1}\mathrm{x}_{i-1,r} + \varepsilon_{ir}, \quad (2)$$

where the error terms ε_{ir} (r = 1,...,n) are independent and identically distributed as a normal distribution with mean zero and variance ν_i^2. When i = 1, the regression model (2) is understood as a location case. The regression parameters in (2) can be expressed in terms of the μ_i and the inverse Cholesky root of Σ. We will briefly derive them for easy reference. Let $A = (a_{ij})$ be a lower triangular matrix with positive diagonal elements such that $\Sigma = AA^T$. Let $B^T = A^{-1}$. Then B is an upper triangular matrix with positive diagonal elements and its (j,i)th element will be denoted by b_{ji}. The Cholesky root A and the inverse Cholesky root B are uniquely determined. Let $\mathrm{z}_{(i)}$ be the i by 1 vector formed by the first i elements of a column vector z. We denote by $\mathrm{Q}_{(i)}$ the leading principal submatrix having order i of a square matrix Q. Then the marginal distribution of $\mathrm{x}_{(i)}$ is a multivariate normal with mean vector $\mu_{(i)}$ and covariance matrix $\Sigma_{(i)}$. The conditional expectation of x_i given $\mathrm{x}_1, \cdots, \mathrm{x}_{i-1}$ is

$$\mu_i + \sigma_{i.21}\Sigma_{(i-1)}^{-1}(\mathrm{x}_{(i-1)} - \mu_{(i-1)}),$$

where $\sigma_{i.21}$ is a row vector of size i-1 consisting of the first i-1 elements in the last row of $\Sigma_{(i)}$ and its residual variance is given by

$$\sigma_{ii} - \sigma_{i.21}\Sigma_{(i-1)}^{-1}\sigma_{i.12},$$

where σ_{ii} is the ith diagonal element of Σ and $\sigma_{i.12} = \sigma_{i.21}^{T}$. By partitioning $\Sigma_{(i)}$ into $\Sigma_{(i-1)}$ and the others in an obvious way, and using the identities $\Sigma_{(i-1)} = A_{(i-1)}A_{(i-1)}^{T}$ and $B_{(i-1)}^{T} = A_{(i-1)}^{-1}$, we easily get the vector of regression coefficients as

$$\sigma_{i.21}\Sigma_{(i-1)}^{-1} = -\alpha_{ii}(\mathbf{b}_{1i},\cdots,\mathbf{b}_{i-1,i}).$$

Since $\sigma_{i.21}\Sigma_{(i-1)}^{-1}\sigma_{i.12} = \Sigma_{k=1}^{i-1}\alpha_{ik}^{2}$, the residual variance becomes $\mathbf{b}_{ii}^{-2}$. Hence we have

$$\beta_{i0} = \mu_i + \Sigma_{k=1}^{i-1}(\mathbf{b}_{ki}/\mathbf{b}_{ii})\mu_k, \qquad \beta_{ik} = -\mathbf{b}_{ki}/\mathbf{b}_{ii} \quad (1 \le k \le i-1) \tag{3}$$

and $\nu_i^2 = \mathbf{b}_{ii}^{-2}$. The identities (3) can be found in Hawkins and Eplett (1982).

For each ordering of the variables in a random vector, we have p regressions. For each i, the least squares estimators for the regression coefficients and the residual variance are the same as those in (3) with the μ_i and $\mathbf{b}_{ki}$ replaced by the counterparts of the sample mean vector and the sample covariance matrix with its divisor equal to the sample size (Johnson and Wichern, 1992, pp.338-340). Thus the identities (3) show that the mean vector and the covariance matrix can be estimated by successively fitting a multivariate normal data to the regression model (2) from i=1 to p. At each step μ_i and the ith column of B are determined. This process of estimation is adopted only for screening out possible outliers and based on the remaining data we will determine whether or not observations screened out are outliers. Observations not classified as outliers can then be used for robust estimation.

Our procedure for detecting outliers is as follows:

1. The least median of squares method is applied to all the regressions for each ordering of the variables and select good observations.
2. Based on the selected observations find the maximum likelihood estimators for μ and Σ.
3. Classify the observations not selected in Step 1 as outliers or not according to the corresponding Mahalanobis distance obtained by replacing $\mathbf{m}$ and $\mathbf{W}$ by the maximum likelihood estimators in Step 2

Let r be the number of observations selected in Step 1 and $\mathbf{x}_0$ be an observation not selected. Then $\mathbf{x}_0 - \mathbf{m}$ is distributed as a p-variate normal with mean zero and

covariance matrix $[(r+1)/r]\Sigma$ and it is independent of **W**. Since $r\mathbf{W}$ is distributed according to a Wishart distribution with scale matrix Σ and degrees of freedom r-1,

$$\frac{r-p}{p(r+1)}(\mathbf{x}_0 - \mathbf{m})^T\mathbf{W}^{-1}(\mathbf{x}_0 - \mathbf{m})$$

has a F-distribution with degrees of freedom p and r-p (Seber, 1984, p.30). Thus observations with sufficiently small p-value are regarded as outliers.

3. Examples

Two data sets are given for illustration. Two cases in which **m** and **W** are the usual sample mean vector and the sample covariance matrix (MD) and the estimators described in Step 2 (RD) are considered and compared. The utility function *lmsreg* in Splus (1990) is used to compute the LMS estimates for the regression parameters. A Fortran program named PROGRESS is also available in the library *statlib* at Carnegie-Mellon University.

Table 1 shows the Mahalanobis distances for the cost data measured on three variables (Johnson and Wichern, 1992, p.217). Bacon-Shone and Fung (1987) and Caroni and Prescott (1992) analyzed the cost data based on the usual Mahalanobis distance. They concluded that observations 9 and 21 are highly significant. After applying the procedure in Section 2, eight observations are not selected at Step 1 and they are listed in Table 2. Taking a look at the p-values for the eight observations, observation 9 is highly significant and 21 is the next. Observations 36 and 20 are a little significant. The others do not seem to be significant.

Hawkins, Bradu and Kass (1984) constructed a data set to illustrate the masking effect in the context of regression analysis. Rousseeuw and van Zomeren (1990) analyzed the data set only for three independent variables using a robust version of the Mahalanobis distance. Their conclusion is that observations 1 to 14 are outliers. Our analysis is also confined to the data for three independent variables. Two types of distances are included in Table 3. The usual Mahalanobis diatances do not reveal the outliers well. Observations 12 and 14 masks the others. The first step in our procedure removes the first 14 observations. The resulting Mahalanobis distances identify the outliers very well. The p-values for the removed observations are nearly zero and therefore the first 14 observations are highly significant.

Table 1. Mahalanobis distances for Cost Data

No	MD	RD	No	MD	RD	No	MD	RD
1	1.10	1.95	13	0.84	0.87	25	2.45	3.16
2	1.81	1.98	14	0.63	0.66	26	1.53	1.65
3	1.86	1.84	15	2.08	2.78	27	1.92	2.57
4	1.83	3.79	16	1.54	1.59	28	1.18	1.81
5	0.98	0.92	17	1.03	1.43	29	1.57	1.52
6	1.17	1.93	18	1.77	2.16	30	1.34	2.07
7	0.72	0.97	19	0.64	0.67	31	1.76	2.54
8	1.81	2.16	20	2.56	4.13	32	1.58	2.63
9	4.26	10.0	21	3.32	5.95	33	1.15	1.64
10	0.34	0.50	22	0.71	1.32	34	1.06	1.94
11	1.03	1.94	23	2.29	2.60	35	1.46	1.58
12	1.13	1.05	24	1.39	1.58	36	2.10	4.50

Table 2. P-values for Cost Data

No	4	9	20	21
p-value	1.66×10^{-2}	2.52×10^{-8}	8.17×10^{-3}	1.45×10^{-4}
No	25	27	32	36
p-value	5.62×10^{-2}	1.55×10^{-1}	1.42×10^{-1}	3.72×10^{-3}

Table 3. Mahalanobis distances for the Hawkins-Bradu-Kass data

No	MD	RD	No	MD	RD	No	MD	RD
1	1.93	29.69	26	1.18	1.72	51	1.31	1.53
2	1.87	30.46	27	1.46	2.01	52	2.09	2.11
3	2.33	32.16	28	0.87	1.04	53	2.23	2.54
4	2.24	33.13	29	0.58	1.15	54	1.42	1.92
5	2.11	32.55	30	1.58	2.13	55	1.24	1.39
6	2.16	30.84	31	1.85	1.73	56	1.34	1.66
7	2.02	30.94	32	1.32	1.78	57	0.84	1.37
8	1.93	30.05	33	0.99	1.29	58	1.41	1.76
9	2.24	32.22	34	1.18	2.06	59	0.60	1.29
10	2.35	31.20	35	1.25	1.90	60	1.90	2.11
11	2.46	36.94	36	0.86	1.15	61	1.69	2.26
12	3.13	38.27	37	1.84	2.03	62	0.76	2.00
13	2.68	37.22	38	0.76	1.50	63	1.30	1.82
14	6.42	41.43	39	1.27	1.81	64	0.98	1.87
15	1.83	2.02	40	1.12	1.09	65	1.16	1.56
16	2.17	2.18	41	1.71	2.06	66	1.31	1.47
17	1.39	1.95	42	1.78	1.81	67	0.63	0.55
18	0.85	0.79	43	1.88	2.18	68	1.56	2.12
19	1.16	1.30	44	1.43	2.04	69	1.08	1.76
20	1.60	2.08	45	1.08	1.87	70	1.00	1.37
21	1.10	1.07	46	1.35	1.89	71	0.65	1.01
22	1.56	1.76	47	1.98	2.30	72	1.06	0.93
23	1.09	1.17	48	1.43	1.90	73	1.48	1.33
24	0.98	1.33	49	1.58	1.60	74	1.66	1.53
25	0.80	2.00	50	0.43	1.49	75	1.91	2.08

References

[1] Bacon-Shone, J. and Fung, W.K. (1987). A New Graphical Method for Detecting Single and Multiple Outliers in Multivariate Data. *Appl. Statist.* 36, 153--162.

[2] Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data*, John Wiley & Sons, New York, 2nd edition.

[3] Campbell, N.A. (1980). Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation. *Appl. Statist.* 29, 231--237.

[4] Caroni, A. and Prescott, P. (1992). Sequential Application of Wilks's Multivariate Outlier Test. *Appl. Statist.* 41, 355--364.

[5] Hawkins, D.M., Bradu, D and Kass, G.V. (1984). Location of Several Outliers in Multiple-Regression Data Using Elemental Sets. *Technometrics* 26, 197--208.

[6] Hawkins, D.M. and Eplett, W.J.R. (1982). The Cholesky Factorization of the Inverse Correlation or Covariance Matrix in Multiple Regression, *Technometrics* 24, 191--197.

[7] Johnson, A.J. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. Englewood Cliffs: Prentice-Hall.

[8] Rousseeuw, P.J. (1984). Least Median of Squares Regression. *J. Amer. Statist. Ass.* 79, 871--880.

[9] Rousseeuw, P.J. and Leroy, A.M.(1987). *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

[10] Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. *J. Amer. Statist. Ass.* 85, 633--651.

[11] Seber, G.A.F. (1984). *Multivariate Observations*, John Wiley & Sons, New York.

[12] S-Plus for Dos : User's Manual, 1990, Statistical Sciences, Inc. Seattle, Washington.

[13] Wilks, S.S.(1963). Multivariate Statistical Outliers. *Sankhyā* 25), 407--426.